

The Role of Synthetic Data in Research: Benefits, Costs, and Practical Insights from Data Owners and Trusted Research Environments Experts

Cristina Magder, Maureen Haaker, Jools Kasmire, Hina Zahid, Melissa Ogwayo



Contents

Executive Summary.....	4
Background and purpose	4
Key findings	4
Recommendations at a glance.....	5
Introduction	7
Work Package 1: Understanding the landscape of synthetic data	9
Purpose and methodology	9
What we discovered	9
Implications for the project.....	10
Work package 2: Survey with data creators	11
What we asked	11
Who responded?	11
Findings: What we learned	12
What this means for the future	13
Work package 3: Case studies with data creators.....	14
How organisations are using synthetic data	14
What it takes to produce synthetic data	14
Key challenges in creating and sharing synthetic data	15
Lessons learned and future opportunities	16
What this means for the broader data community	16
Work package 4: Focus group with TRE representatives	17
Why this focus group was important	17
What we learned.....	17
Key challenges in implementing synthetic data in TREs	18
What this means for the future of synthetic data in TREs.....	19
Discussion: Bringing the findings together	21
Limitations of the project	22
Next steps: unlocking the full potential of synthetic data.....	23
Recommendations	24

Final reflections.....	27
Acknowledgements.....	27

Executive Summary

Background and purpose

Access to high-quality, detailed, sensitive and secure data is essential for research, policymaking, and innovation, but it is not always straightforward. Strict legal and ethical safeguards are in place to protect research participants and ensure that data are used responsibly. Due to these necessary safeguards, researchers must complete several steps before they can begin their work, often including securing accreditation, writing detailed proposals, and selecting the specific variables required for their analyses. These safeguards are essential, but they can also slow down research, particularly when demand for data outstrips available resources.

Trusted Research Environments (TREs) provide secure access to detailed, sensitive data, but managing these data requests, including reviewing proposals and ensuring compliance of people, project, environments and outputs, takes time and resources. Balancing efficiency with legal and ethical standards is a constant challenge. Synthetic data offers one possible way to ease some of this pressure. For example, it can allow researchers to explore data structures before applying for access. This is intended to help reduce unnecessary requests, empowering researchers to write more effective applications and streamline existing processes without compromising essential safeguards.

This project was funded to explore low-fidelity synthetic data's role in improving data access to secure and sensitive data, particularly from the perspective of data creators, data owners, and TRE representatives. Complementary projects have focused on the needs of researchers and the public views.

We focused on three key areas:

- What does it cost, financially and operationally, to produce and maintain synthetic data?
- What governance and sharing models have been explored and work best for synthetic data?
- How could synthetic data help improve research workflows and reduce the burden on TREs?

To explore these questions, we began our work by conducting a literature review, followed by surveys of data creators, case studies with key organisations already piloting synthetic data, and a focus group with representatives from TREs.

Key findings

Synthetic data has clear potential, but adoption remains slow.

Many organisations recognise that synthetic data can help improve access to secure and

sensitive data. However, only a small number are actively using it. The lack of standardisation, concerns about quality, and governance uncertainties are key barriers to broader adoption.

Synthetic data can make research more efficient.

Data owners and TRE professionals recognise that synthetic data enables researchers to explore data structures before applying for real data, speeding up approvals, reducing request errors, and easing the workload on TREs. Organisations such as NHS England and the Office for National Statistics (ONS) are already using and exploring synthetic data in this way.

Costs and resource constraints are a significant challenge.

Generating high-quality synthetic data requires specialist skills, computing resources, and ongoing maintenance. Overall, while synthetic data is considered cost-effective, some organisations still struggle with the upfront investment. However, all case studies revealed that those who have piloted synthetic data found it both beneficial and worthwhile. The challenge lies in central budgets, which currently do not prioritise synthetic data.

Lack of quality assurance standardisation is limiting progress.

Different organisations use different methods to create, validate, and share synthetic data. The absence of common licensing agreements, quality assurance checks, and protocols for sharing synthetic data makes it more challenging for data creators and TREs to integrate it into their processes with confidence. Without a standard framework, data owners and TREs struggle to determine when, how, and under what conditions synthetic data should be created and shared.

Unclear governance frameworks and legal guidance create uncertainty.

There is no clear legal position on whether synthetic data falls under UK GDPR or how it should be governed. Some data creators treat synthetic data as low-risk and make it available under Open Licences or standard terms. In contrast, others apply stricter access controls, leading to inconsistencies in the landscape. Without clear governance, the adoption of synthetic data is likely to remain inconsistent and uncoordinated.

Recommendations at a glance

Policymakers: Clarify governance and legal frameworks.

Policymakers should provide clear guidance on how synthetic data fits within existing privacy laws, data-sharing policies, and risk management frameworks. This will help data creators and TREs make informed decisions about its use, licensing, and access controls.

Funding bodies: Strengthen skills and infrastructure.

Organisations need sustained investment in funding, training, and computational resources to develop and maintain high-quality synthetic data. Providing targeted support will help organisations build the skills and resources needed to produce, validate, and integrate synthetic data effectively.

Data creators and TREs: Establish quality standards and sharing models.

A consistent framework for generating, evaluating, and sharing synthetic data is needed. Developing standardised quality assurance checks, licensing agreements, and validation methods will help organisations trust and use synthetic data with confidence.

The research community: Improve awareness, training and public engagement.

Building trust in synthetic data requires a coordinated effort across the research community, including funding bodies, policymakers, data creators, data providers, TREs, and the public. By working together, the research community can ensure that synthetic data is used responsibly, transparently, and effectively to support innovation while maintaining privacy and trust.

This work was supported by [ADR UK](#) (Administrative Data Research UK). ADR UK is a partnership transforming the way researchers access the UK's wealth of public sector data, enabling better-informed policy decisions that improve people's lives. ADR UK is an Economic and Social Research Council (ESRC) investment (part of UK Research and Innovation).

Grant number: ES/Z502467/1

Introduction

Granular data is essential for research, innovation, and policymaking. However, using secure and sensitive detailed data presents challenges, including restrictions in line with ethical and legal frameworks, and often necessitates a multi-step process before access can be granted. One possible solution to alleviate these challenges is the use of synthetic data.

The existing literature reveals a spectrum of synthetic data types, ranging from low-fidelity representations used for basic structural descriptions of data to high-fidelity datasets that closely mimic real data. Given these variations, establishing a clear and consistent definition of synthetic data was essential for this study.

For the purposes of this project, synthetic data refers to artificially generated data. These data can either be created using real data or can be “data-free”, created by using only metadata that includes structural information, such as data types, formats, and summary statistics. Synthetic data was categorised into two distinct types:

- Low-fidelity synthetic data: Datasets that maintain structural elements of real data but lack detailed statistical relationships between data elements. These are suitable for initial exploratory research and familiarisation with data.
- High-fidelity synthetic data: Datasets that closely replicate the statistical distributions and interdependencies of the original dataset while mitigating privacy risks.

While both types serve distinct purposes, this project primarily focused on low-fidelity synthetic data, assessing its role in improving access to secure data. However, high-fidelity synthetic data was also considered where relevant, particularly in the survey and case studies, as some organisations are exploring its use. Including both perspectives provided a more comprehensive understanding of synthetic data adoption, governance, and potential applications across various settings.

To explore these issues, we collaborated closely with data owners and TRE representatives, assessing whether synthetic data could enhance efficiency, streamline data access, and alleviate operational burdens. The project focused on three key areas:

- What are the financial and operational costs of producing and maintaining synthetic data?
- How can synthetic data be shared practically, securely, and responsibly?
- Does access to synthetic data help researchers work more effectively and ease the burden on secure data environments?

To answer these questions, we reviewed existing literature and surveyed data owners. We also carried out case studies with organisations already working with synthetic data and spoke to TRE experts to understand how synthetic data fits into the broader data landscape.

Work Package	Aim	Method
Work Package 1: Understanding the landscape of synthetic data	To map existing research, practices, and frameworks related to synthetic data, identifying gaps and opportunities.	Literature review of published studies, reports, and operating procedures.
Work Package 2: Survey with Data Creators	To understand how synthetic data is being created and used, challenges faced, and future expectations from data creators.	Online survey distributed to government bodies, universities, NHS, and non-profits.
Work Package 3: Case Studies with Data Creators	To explore practical realities of synthetic data use within specific organisations.	Case studies conducted with NHS England, ONS, MoJ, and DfE.
Work Package 4: Focus Group with TRE Representatives	To gather insights from TREs about the role of synthetic data in secure data access.	Focus group discussion with representatives from various TREs.

This report presents our findings and practical recommendations for policymakers, funders, and data creators and the wider research community.

Work Package 1: Understanding the landscape of synthetic data

Before exploring synthetic data's practicalities with data owners, we needed to understand the existing research. This involved reviewing studies, reports, and documented best practices to map what is already known, identify gaps, and highlight relevant lessons for this project. The literature review was conducted using a thematic approach, synthesising findings from peer-reviewed articles, institutional reports, grey literature, and existing openly available guidance from key data owners such as the Office for National Statistics, the NHS, and the Ministry of Justice.

Purpose and methodology

The purpose of this literature review was to explore the current landscape of synthetic data generation, governance frameworks, applications, challenges, and opportunities. This review aimed to identify existing methodologies, tools, ethical and legal considerations, and potential future trends relevant to the use of synthetic data across different domains.

Searches were carried out across several academic databases including Web of Science, Scopus, and Google Scholar, using keywords such as “synthetic data generation,” “ethical frameworks,” “tools for synthetic data,” and “costing models for synthetic data.” Additionally, targeted searches were conducted to gather information from relevant grey literature and stakeholder reports.

Due to the broad scope of the search and the time constraints, a convenience sampling strategy was employed, focusing on articles that provided comprehensive overviews or frameworks relevant to synthetic data generation. A total of 50 articles were selected based on their relevance to the objectives of this review. Furthermore, snowballing techniques were used to identify additional sources from reference lists of key articles.

What we discovered

Synthetic data creation

The creation of synthetic data can differ from simple statistical methods to advanced artificial intelligence techniques. Some methods focus on privacy, ensuring that no real-world data can be traced back, while others prioritise realism, making the data as close to the real thing as possible. The choice of approach depends on the purpose, whether it's for testing, code development, training, or specific analysis.

Ethical and legal considerations

The legal and ethical landscape for synthetic data is still evolving. While it offers clear advantages for privacy protection, it isn't fully understood what, if any, process fall under data protection legislation such as the UK GDPR.

Questions remain about who owns synthetic data and whether it should be subject to the same governance as real data. It is unclear whether ownership automatically remains with the original data owner, especially when the synthetic data is generated using only metadata or derived models, or whether the creator of the synthetic dataset has a claim to ownership.

This ambiguity complicates licensing decisions and responsibilities for onward use, particularly in collaborative or cross-organisational settings. Organisations like the Office for National Statistics (ONS), NHS England, and the Ministry of Justice are piloting synthetic data in the UK. Still, there is no standardised approach to its regulation or use.

Cost considerations

Another area we explored was cost. Synthetic data is often described as a cost-saving tool because it can reduce operational challenges, streamline access process and even sometimes reduce the need for direct access to secure and sensitive data. However, the actual financial impact is not well understood. Producing high-quality synthetic data requires investment in specialised software, skilled personnel, and computing power. There's little research on how these costs compare to the savings generated from improved data access and efficiency.

Challenges and future directions

We also looked at emerging challenges and future directions. While synthetic data is gaining traction, concerns about its quality, how it should be evaluated, and how to prevent misuse exist. Evaluation, in this context, goes beyond simple quality assurance. It includes checking statistical validity, assessing fitness for purpose, and reviewing the processes used to generate synthetic data. This is particularly important when personal data are used to create synthetic data. In this instance, the legal basis for processing the original personal data must be clearly established, including whether it can be used for the creation of the synthetic data. Public understanding of synthetic data is limited, and there are fears that it could be mistaken for real data, leading to unintended consequences. As the technology continues to evolve, there is a growing need for more precise standards, better validation methods, and stronger governance frameworks.

Implications for the project

This review helped us see the bigger picture. It highlighted the opportunities that synthetic data presents and the practical challenges that need to be addressed. It gave us a solid foundation for the next stage of our research, where we engaged directly with data creators, data owners and TRE representatives to understand their experiences.

Work package 2: Survey with data creators

To understand how synthetic data is used in practice, we designed a survey for organisations involved in data creation, management, and governance. Our goal was to capture a snapshot of where synthetic data fits into their work, their challenges, and their views on its future.

The survey was distributed using a mixed recruitment strategy. Targeted emails were sent to selected organisations known to be involved in the creation and management of secure and sensitive data, including government departments, healthcare bodies, and research institutions. In addition, the survey was shared more broadly via professional mailing lists and social media platforms to encourage wider engagement. While open in distribution, the framing of the survey clearly identified data creators, data owners, and those managing data access that can respond on behalf of their organisations as the intended respondents. This approach allowed us to capture a diverse but relevant sample of stakeholders actively working at the intersection of data security, access, and innovation.

What we asked

To get a complete picture of synthetic data use, we designed the survey around key questions that explored different aspects of data creation and management:

- Current use of synthetic data – Are organisations already generating synthetic data, or are they still exploring its potential? If they are using it, what kind of synthetic data are they creating and what methods are they relying on?
- Motivations and barriers – What are the main reasons organisations are interested in synthetic data? What concerns or challenges are preventing them from adopting it more widely?
- Technical and financial considerations – What tools and processes are used to create synthetic data? How much does it cost to produce, and where does the funding come from?
- Data sharing and governance – How is synthetic data being shared, and what safeguards are in place to ensure responsible use? Are there concerns about security, misuse, or public perception?
- Future plans – Are organisations planning to expand their use of synthetic data? What support or guidance would help them do so more effectively?

Who responded?

The synthetic data production survey received fifteen responses from various organisations, including the central government, universities, the NHS, non-departmental public bodies, and the voluntary sector. Most responses came from government departments and higher education

institutions, reflecting strong interest from the public sector, but contributions from healthcare and non-profits show that synthetic data is gaining attention across different fields.

Although synthetic data is not a new concept, its use as a tool in data access infrastructure is still emerging, particularly in terms of developing consistent standards, governance frameworks, and clear definitions. While some organisations have been exploring it for over five years, the majority of the respondents were still in the early stages of engagement, with 60% saying they had only started assessing its potential in the last two years. This highlights that, while awareness is increasing, there is still a long way to go before synthetic data becomes a mainstream solution.

Findings: What we learned

Interest is growing, but adoption is slow.

There is a clear interest in synthetic data, particularly improving access to secure and sensitive data. However, only a small number of organisations are actively using it. Out of the fifteen survey respondents, just three reported actively producing synthetic data and one generating “data-free synthetic data” based solely on metadata. Two organisations were in the planning or assessment phase, and three were considering future production. The remaining six had no current or short-term plans to produce synthetic or data-free synthetic data. These figures highlight that while interest exists, concerns around technical feasibility, governance, and the lack of clear guidance are still slowing down broader adoption.

The main drivers: privacy, efficiency, and research access

Organisations see synthetic data as a practical way to balance the need for open research with privacy and security requirements. One of its most significant advantages is that it allows researchers to explore datasets before applying for access to the real data. This can speed up the approval process, reduce the workload for data providers, and make research more efficient, as well as reduce the need to provide access to real data when it isn’t necessary – this was especially relevant for the twelve respondents who produce research data from which eleven share real data through TREs. Synthetic data was also recognised as useful for training AI models, testing software, and enabling innovation in data science.

Cost and resources are significant barriers.

One of the biggest challenges facing organisations is the cost of synthetic data production. While some described it as a cost-effective alternative to managing secure access to real data, others highlighted that generating high-quality synthetic data requires significant investment in specialist staff, software, and computing power. Reported costs varied widely, with some organisations estimating annual budgets of up to £250,000, while others worked with much smaller amounts. Most relied on general research budgets, rather than dedicated funded, making prioritising investment in this area difficult. Survey responses also reflected a range of technical approaches to synthetic data generation: two organisations reported using Python-based tools, two relied on custom-built solutions, and one used Oracle systems. Regardless of

the tool, respondents consistently emphasised that significant internal resources were required to produce and maintain synthetic datasets.

Concerns about data quality and potential misuse

Another primary concern for organisations is ensuring that synthetic data is of high enough quality to be useful while avoiding the risk of misinterpretation. If synthetic data is not well-designed and well documented, it could mislead researchers. There is uncertainty about how synthetic data should be labelled and documented, with fears that people might mistake it for real data, leading to incorrect assumptions or inappropriate use with governance and oversight reflecting as key concerns.

No standard approach to sharing and licensing

While some organisations make synthetic data openly available, others only allow access through secure environments. Some organisations have introduced safeguards such as specific terms of use for synthetic data, while others treat it the same way as real data regarding access requirements. Among survey respondents, only one currently shares low fidelity synthetic data openly, and two others do so under specific conditions. This means that researchers face different rules and restrictions depending on where the data are hosted.

A need for more explicit guidance and support

Many organisations highlighted that they would benefit from more explicit guidance on best practices for synthetic data. In particular, they would like support in:

- Developing standardised quality checks and evaluation frameworks;
- Creating governance and security protocols that are appropriate for different types of synthetic data; and
- Improving awareness and training for researchers so they understand how to use synthetic data effectively and responsibly.

What this means for the future

The survey confirmed that synthetic data is a promising tool, but uncertainty, cost barriers, and a lack of standardisation are holding back its adoption. Many organisations are interested but unsure about the best way to proceed. Without more precise guidance and dedicated funding, progress will likely remain slow.

These insights have helped shaped the next stage of our research, where we carried out case studies to explore how different organisations tackle these issues in practice. By looking at real-world examples, we aimed to identify practical solutions and best practices that could help unlock the full potential of synthetic data.

Work package 3: Case studies with data creators

To examine how organisations are producing and sharing synthetic data we conducted a series of case studies with key organisations that are actively working in this area. The aim was to understand the practical realities, what works, what doesn't, and what lessons can be learned for the wider data community.

These organisations were identified during the bid development phase, based on their known involvement in synthetic data initiatives at the time. We focused on four organisations that represent different approaches to synthetic data:

- NHS England – developing synthetic versions of health data for research and training.
- Ministry of Justice (MoJ) – using synthetic data to support access to justice datasets.
- Office for National Statistics (ONS) – exploring the use of synthetic data within the Integrated Data Service (IDS).
- Department for Education (DfE) – exploring synthetic versions of student data, with insights gathered from an interview with a researcher working in collaboration with DfE.

Each organisation is at a different stage in its synthetic data journey, but all have faced similar challenges and opportunities.

How organisations are using synthetic data

One of the most common uses of synthetic data is to allow researchers to explore datasets before applying for access to the real data. This “preview” function helps researchers refine their data requests, reducing delays and making the access process more efficient for both researchers and TREs. For example, the ONS has developed synthetic versions of population data, allowing researchers to test their code and methods before accessing the real data.

In some cases, synthetic data is also used for training and education. The NHS England Hospital Episode Statistics (HES) synthetic dataset has been used in university courses, allowing students to practise working with health data. This has been considered a key achievement for those involved as it helps build skills without the risks associated with using real patient records.

Another emerging use case is machine learning and AI training. Synthetic data allows testing and developing algorithms without the need to use sensitive information. The MoJ is exploring how synthetic versions of court and prison datasets can be used in predictive modelling while reducing privacy risks.

What it takes to produce synthetic data

Creating a useful synthetic dataset is not as simple as pressing a button. The process involves multiple steps:

1. Defining the purpose: what will the synthetic data be used for? Does it need to be high-fidelity (closely resembling real data) or low fidelity (just capturing the basic structure)?
2. Choosing the right methods: depending on the complexity of the data, techniques range from simple statistical models to advanced machine learning.
3. Generating and testing the data: this involves checking that the synthetic data behaves realistically without being too similar to real records.
4. Documentation and governance: organisations must clearly label synthetic datasets and provide guidance on how they should be used.

Each organisation in our study took a slightly different approach depending on its needs and resources. Some relied on established statistical methods, while others experimented with artificial intelligence techniques to create more realistic synthetic datasets.

The biggest challenge, however, was quality assurance. Organisations had to ensure that the synthetic data was helpful in research without being misleading or too similar to real data. ONS, for example, applies a series of statistical checks to verify that its synthetic data provide a reliable representation of trends without risking privacy concerns.

Key challenges in creating and sharing synthetic data

Time and resources

Generating high-quality synthetic data takes time. Across all organisations, one of the most significant barriers was staffing and expertise. Producing a well-documented dataset can take several months and requires skilled data scientists and governance teams. Some organisations estimated that generating a synthetic dataset could take between three to six months of full-time work by a data scientist.

Costs and sustainability

There is no standard pricing model for synthetic data production. Some organisations found it cost-effective because it reduced the need for complex data access processes, while others saw it as an additional cost for managing real data. The most expensive parts of the process included hiring or training staff with expertise in synthetic data generation, computational resources and software and the ongoing updates and maintenance of synthetic data collections. Some organisations struggled to justify continued pursuance in synthetic data projects without dedicated funding.

Governance and public perception

While synthetic data is often seen as a privacy-enhancing tool, governance hurdles still exist to overcome. A common challenge was getting internal approval for synthetic data projects, as many decision-makers were unsure whether synthetic data should be subject to the same legal and ethical scrutiny as real data.

Public understanding of synthetic data also varied. Some organisations are worried that users might misinterpret synthetic data as real data or, conversely, assume it is too artificial to be

helpful. This has led some organisations, such as DfE, to require that synthetic datasets come with clear documentation and disclaimers outlining their intended use.

Lack of standardisation

Each organisation had its way of creating and sharing synthetic data, leading to inconsistencies across sectors. Some made synthetic data available for download under Open or bespoke licences, while others restricted access to secure research environments. There is currently no agreed-upon best practice for labelling, licensing, or distributing synthetic data, which makes it harder for researchers to use synthetic datasets consistently.

Lessons learned and future opportunities

Despite the challenges, organisations in our case studies agreed that synthetic data has the potential to play a more significant role in data access and research. The key lessons learned include:

- Synthetic data should be designed with a clear purpose: low-fidelity synthetic data works well for testing and training, while high-fidelity synthetic data is more suitable for detailed analysis.
- Governance frameworks need to be more explicit: organisations need better guidance on managing, documenting, and sharing synthetic data responsibly.
- Standardisation would help adoption: a set of standard practices for quality assurance, licensing, and metadata would make synthetic data more straightforward to use across different research environments.
- More investment is needed in training and skills: many organisations lack in-house expertise, meaning that more training programmes or knowledge-sharing networks could help drive adoption.

What this means for the broader data community

While synthetic data is not a replacement for real data, it has the potential to complement secure data access and make research faster and more flexible.

The findings from this work package fed into our final stage of research, where we engaged with TRE representatives to explore how synthetic data fits into the broader ecosystem of secure data access.

Work package 4: Focus group with TRE representatives

As the final stage of our research, we held a focus group with representatives from TREs to explore the role of synthetic data in secure data access. TREs ensure that sensitive data is handled safely, providing controlled access to researchers to secure and sensitive data such as health records and government data in line with ethical and legal standards and requirements.

This discussion was conducted under Chatham House Rules, meaning participants could speak openly without being directly quoted and identified. This allowed for an open exchange of views, helping us capture the real concerns, experiences, and priorities of those working in secure data environments. Participants were recruited through targeted outreach to known TRE networks included the ADR UK synthetic data working group, the Safe Data Professional Network as well as the UK TRE mailing list. The session included TREs from six national and institutionally hosted TREs.

Why this focus group was important

In the UK several TREs are already exploring synthetic data, but there is no consistent approach to its use. Some see it as a valuable tool to improve research efficiency, while others have concerns about governance, legal risks, and public perception.

The focus group aimed to:

- Understand how synthetic data is currently being used within TREs;
- Identify the main challenges and concerns around its adoption; and
- Explore potential governance models to ensure safe and responsible use.

What we learned

There is still a lack of consensus around what synthetic data is.

While synthetic data is widely discussed in the research landscape, there is still some confusion about how it differs from anonymised or de-identified data. Some TRE representatives noted that researchers and policymakers often assume that synthetic data is just a modified version of real data rather than an entirely computer generated dataset. This confusion can create resistance to adoption, particularly when securing legal or ethical approval for synthetic data projects.

Synthetic data could improve efficiency - but only if it's trusted.

One of the most significant potential benefits of synthetic data is that it could speed up research by allowing users to familiarise themselves with a dataset before applying for real data access. This could reduce the number of incorrect or overly broad data requests, making the approval process more efficient.

However, TRE representatives that participated in our focus group agreed that researchers need to trust synthetic data for this to work. Synthetic datasets that are poorly constructed, missing key variables, or too different from real data will not be useful. Some participants suggested that synthetic data should always come with clear documentation explaining its limitations so that researchers understand what it can and cannot be used for.

Governance remains a grey area.

There is currently no standard governance framework for synthetic data in TREs. Different organisations take different approaches, with some treating synthetic data as a low-risk resource while others apply the same strict access controls as they do for real data.

The main concerns raised included:

- Legal uncertainty – No clear legal guidance on whether synthetic data falls under data protection laws like GDPR. Some organisations worry that even though synthetic data is not real, it could still be considered personal data if it accurately reflects patterns in the original dataset;
- Risk of misinterpretation—If synthetic data is too realistic, users might mistake it for real data and draw incorrect conclusions; and,
- Ethical concerns – Some participants argued that while synthetic data reduces privacy risks, it still requires careful governance to prevent misuse or unintended consequences.

Key challenges in implementing synthetic data in TREs

Public perception and trust

A recurring theme in the discussion was how the public might perceive synthetic data. While data professionals see it as a privacy-friendly solution, members of the public may be sceptical or even distrustful of its use. Some TREs reported concerns that synthetic data might be seen as a way to "bypass" traditional safeguards or make sensitive data more widely available than it should be.

There is also the risk of misunderstanding synthetic data's accuracy. If researchers assume synthetic data is fully representative of real data, they might use it for analysis it was never intended for. TRE representatives agreed that stronger communication and guidance are needed to clarify synthetic data, how it should be used, and what safeguards are in place.

Balancing access and security

Some TREs are open to making synthetic data more widely available, while others prefer to restrict access in the same way as real data. There was no clear consensus on the best approach, but some TREs suggested that different tiers of synthetic data could be created:

- Low-fidelity synthetic data – Could be openly available to help researchers explore dataset structures without revealing detailed relationships.

- High-fidelity synthetic data – Would require stricter access controls to prevent potential privacy risks.

One TRE representative explained that, in their view, even synthetic datasets generated from metadata require in-depth privacy assessments to ensure they do not unintentionally reveal patterns that could lead to re-identification. This raises questions about whether re-identification is always the appropriate term in this context and highlights the broader challenge of defining synthetic data and related processes. This suggests that additional and more precise terminology may be necessary to capture the nuances of these concerns.

Need for standardised quality checks

Currently, there are no standardised methods for evaluating the quality and reliability of synthetic data. Some TREs test their synthetic datasets against real data to check for consistency, while others use more informal validation processes.

Participants suggested that creating a set of best practices for synthetic data evaluation would help TREs feel more confident about its use. This could include:

- Clear documentation outlining how the data was generated and its intended purpose;
- Validation metrics to measure how well synthetic data represents key features of real datasets; and
- Standard licensing agreements to ensure consistent rules for how synthetic data can be accessed and shared.

What this means for the future of synthetic data in TREs

The focus group confirmed that while synthetic data has clear potential, there are still barriers to its adoption within TREs. Participants agreed on the following priorities for the future:

- Greater clarity on legal and ethical issues: more guidance is needed on whether synthetic data is subject to existing data protection laws and what governance frameworks should apply.
- More investment in public engagement: to build trust, there should be more transparent communication about how synthetic data is used and what protections are in place.
- Standardisation across TREs: a more consistent approach to quality assurance, licensing, and access controls would make synthetic data more straightforward to use across different research environments.
- Training and upskilling: many TRE teams lack in-house expertise in synthetic data. More training opportunities and knowledge-sharing initiatives could help bridge this gap.

This focus group reinforced that synthetic data is not a magic solution to data access challenges, but it does offer a promising way to improve efficiency. However, for it to be successful, clearer guidance, stronger governance, and better public engagement are needed.

These insights fed into our final recommendations, focusing on how policymakers, funders, and data creators can support synthetic data's responsible and effective use across different sectors.

Discussion: Bringing the findings together

Our research has highlighted both the potential and the challenges of using synthetic data to support data access, privacy protection, and research efficiency. Several key themes emerged across the literature, survey responses, case studies, and focus group discussions that provide insight into how synthetic data is currently being used, where the gaps are, and what needs to happen next.

The promise of synthetic data: efficiency, privacy, and accessibility

One of the most substantial findings from this project is that synthetic data is widely seen as a tool for improving access to sensitive datasets while reducing privacy risks. By allowing researchers to explore dataset structures before applying for access, synthetic data can speed up approvals, reduce errors in data requests, and ease the burden on TREs.

Case studies showed that organisations like NHS England and the ONS already use synthetic data, allowing researchers to prepare analysis workflows and test models before working with real data. The focus group discussion confirmed that many TREs recognise these benefits and see synthetic data as a possible solution to the increasing demand for secure data access.

The cost and resource challenge

Despite these advantages, the project found that synthetic data production requires significant time, expertise, and financial investment. The survey results revealed that while some organisations consider synthetic data cost-effective in the long run, others struggle to justify upfront costs. Many reported that hiring skilled staff, acquiring computational resources, and maintaining synthetic datasets require sustained funding, which is not always available. As a result, many organisations face challenges in progressing beyond the planning phase, with resource constraints preventing them from moving forward with implementation.

The case studies provided practical insights into these costs, with organisations estimating that generating a synthetic dataset can take three to six months of full-time work by a data scientist. Focus group participants echoed these concerns, noting that synthetic data production is likely to remain a low priority without dedicated funding streams compared to other data access initiatives.

Governance and public perception: uncertainty remains

A significant barrier to broader adoption is uncertainty around governance and legal frameworks. The focus group discussion highlighted that there is no clear consensus on how synthetic data should be classified—should it be treated as a low-risk resource, or should it be subject to the same controls as real data? Some organisations restrict synthetic data access within TREs, while others provide open access, leading to inconsistency across the research community.

Public perception is another key challenge. The focus group confirmed concerns that researchers and policymakers often misunderstand synthetic data, assuming it is anonymised data rather than an entirely new dataset generated from models. There is also a risk that

synthetic data could be mistaken for real data, leading to misinterpretation. This highlights the need for stronger communication, clear labelling, and standardised documentation to build trust in its use.

Standardisation and best practices: a way forward

Across all work packages, a consistent theme was the lack of standardised methods for evaluating, sharing, and governing synthetic data. Case study participants explained that while they use internal quality checks, there are no widely agreed-upon metrics for assessing synthetic data quality or disclosure risk. The focus group also confirmed that TREs would benefit from guidance on best practices, including:

- Standardised evaluation frameworks to measure how well synthetic data represents real data without compromising privacy
- Clear legal guidance on how synthetic data fits within existing data protection laws
- Common governance models for access, licensing, and security to reduce inconsistencies across different TREs and research institutions

Several participants suggested that a tiered approach to synthetic data access—with low-fidelity synthetic data being openly available and high-fidelity data requiring more restrictions—could help balance accessibility with security. However, consistent rules and governance structures would need to be in place to make this work.

Other emerging opportunities and considerations

While our data did not explicitly explore plug-and-play tools or AI-driven synthetic data generation, these technologies are emerging areas of interest in the wider landscape. Participants did highlight the need for more scalable and sustainable solutions, suggesting that the development of modular, lower-barrier tools could support broader adoption. However, any move toward AI-generated synthetic data must be approached cautiously, with careful consideration of ethical implications, transparency, and fitness for purpose. As the field evolves, these technical innovations may help address some of the cost and resource challenges identified across this study.

Limitations of the project

Although this research provides valuable insights into synthetic data's benefits, costs, and governance challenges, a few limitations must be considered. First, the literature review was conducted within tight timeframes and did not adopt fully systematic methods, which may have led to the omission of new or niche perspectives. Meanwhile, our survey sample size, fifteen organisations, offered variety but may not reflect the full range of data creators or organisations less aware of synthetic data's potential.

The case study organisations were deliberately chosen because they were already active in piloting or implementing synthetic data, so the examples may lean toward more advanced or

successful use cases. Furthermore, the data gathering took place in the UK context, so any recommendations around governance, funding, and infrastructure may not directly transfer to other countries. Adding to that, most participants had a pre-existing interest in synthetic data, which introduces a risk of self-selection bias.

In addition, although we regularly heard concerns about public understanding of synthetic data, we did not directly involve broad-based public engagement. As a result, our insights into public perception come primarily from the viewpoints of data owners and TRE representatives. For a thorough exploration of public perception and recommendations, please see our sister project [DELIMIT](#) “Deliberative workshops with public members: Establishing trust in the use of synthetic data”.

We also discovered a high degree of variability in how organisations track and report costs, preventing us from identifying exact benchmarks for resource requirements. Finally, while our focus centred largely on low-fidelity and some high-fidelity synthetic data, we did not fully explore the rapid rise of AI-driven generation tools, which may pose different benefits, risks, and governance needs. Taken together, these factors should be kept in mind when interpreting the findings and applying them more widely.

Next steps: unlocking the full potential of synthetic data

Building on the findings presented and acknowledging the limitations outlined above, the next steps for advancing synthetic data in research and data access centre on addressing four key barriers:

- **Clear governance and legal frameworks** where policymakers must clarify whether synthetic data falls under existing data protection laws and specify how it should be regulated. Inconsistencies across organisations, some classifying synthetic data as low-risk while others treating it as on par with real data, underscore the urgency for a consistent legal and governance framework.
- **Investment in skills and infrastructure** as many organisations struggle to produce high-quality synthetic data at scale without dedicated funding and expertise. Strengthening budgets for staffing, computational resources, and training will help them move beyond initial planning phases and integrate synthetic data more effectively into their workflows.
- **Standardising quality checks and sharing models and taking** a more consistent approach to evaluating and sharing synthetic data would give researchers, TREs, and data providers greater confidence. Developing shared criteria for data quality, risk assessment, licensing terms, and metadata standards can reduce duplication and enable smoother interoperability between organisations.
- **Better public engagement and researcher training** to address misconceptions about synthetic data, from conflating it with anonymised data to overstating its realism, which

can undermine trust and lead to misuse. Proactive outreach, clearer labelling, and formal training programmes for both researchers and the public will foster more informed, responsible adoption.

Addressing these four areas will require collaboration among policymakers, funders, data creators, and the research community. These next steps can help ensure synthetic data becomes a scalable and trusted resource in the UK research infrastructure and beyond.

Recommendations

This project highlighted that while synthetic data holds clear potential to improve data access and research workflows, several key barriers prevent it from being adopted at scale. These barriers are not technical alone, they span governance, legal uncertainty, funding, consistency of practice, and trust. Based on our findings, we present four detailed, sequenced recommendations. These are structured to reflect what needs to happen first and what depends on it, forming a practical roadmap for responsible scaling of synthetic data across the UK research ecosystem.

1. Policymakers: Clarify governance and legal frameworks

Across the survey, case studies, and focus group, organisations repeatedly flagged legal ambiguity as a critical issue. This included uncertainty over whether synthetic data generation is considered personal data processing under UK GDPR, which lawful basis applies for its creation, who owns the synthetic and how licensing should be approached when synthetic data is shared externally.

These questions create risk aversion and hesitation. Some TREs and data providers treat synthetic data conservatively, applying the same governance standards as they would for real data, while others see it as outside the scope of traditional data protection frameworks. This inconsistency hinders adoption, confuses researchers, and restricts the strategic development of synthetic data.

We recommend that policymakers and regulators collaborate with data controllers, data owners, statisticians, and legal experts to develop clear and practical guidance on the legal and governance aspects of synthetic data. This should include use-case examples, licensing guidance, and clarification of whether different levels of data fidelity entail different legal considerations.

2. Funding bodies: Strengthen skills and infrastructure

Our findings consistently showed that while there is an appetite to explore synthetic data, most organisations lack the sustained capacity to do so. In particular, organisations highlighted difficulty securing dedicated staff time for synthetic data generation, QA, or governance, reliance on general research budgets rather than designated funding streams, a lack of internal expertise in synthetic data generation tools or validation techniques and barriers to accessing or affording the computational resources needed for generation.

This means that even where governance is understood, organisations may still struggle to implement synthetic data workflows. Several participants stated that synthetic data was viewed as an “extra” or “nice to have” rather than a core activity, which risks stalling its development, even when there is a strong strategic interest.

We recommend that funders support long-term investment in the skills, roles, and infrastructure needed to integrate synthetic data into organisational workflows. This could include training programmes for data scientists, analysts, and TRE managers, support for dedicated synthetic data specialist roles or cross-organisational secondments, infrastructure investment (e.g. secure computing environments and software licenses) and provide funding or capacity grants to support pilot work, especially in smaller organisations.

3. Data creators and TREs: Establish quality standards and sharing models

Another key finding was the lack of consistency across synthetic data practice. Some organisations have developed internal protocols for evaluating, documenting, and sharing synthetic data, while others are still deciding whether synthetic data should be treated as “real data” or handled separately. The lack of shared language, tools, and confidence leads to confusion and inefficiencies, especially for researchers moving between different data providers and TREs.

This inconsistency also presents reputational risk. Poorly labelled or low-quality synthetic data can undermine trust in the concept and create the potential for misuse if misunderstood. While standardisation should not mean uniformity, and some flexibility will always be needed, a shared starting point would help TREs and data creators apply consistent practice and reduce duplication.

We recommend that data creators and TREs actively develop and adopt shared models and quality standards for synthetic data. This entails establishing a quality checklist that rigorously assesses fidelity, structure, and utility and adopting clear, standardised documentation and metadata guidelines that transparently explain how each dataset is generated and should be used. We further recommend using harmonised disclaimers and labels to unambiguously indicate the synthetic nature and any limitations of the data.

4. The research community: Improve awareness, training and public engagement

A recurring theme in the focus group and case studies was the risk of misunderstanding synthetic data, both by researchers and the wider public. Examples included researchers using synthetic data as if it were real, misinterpreting synthetic outputs in downstream analysis, or lacking the skills to assess its appropriateness for a given task. There were also concerns that synthetic data might be seen by the public as a way of “bypassing” safeguards, particularly if its role is not clearly explained.

Currently, there are few resources available to help researchers understand how to work with synthetic data responsibly. Meanwhile, public-facing explanations are limited, and engagement strategies are often reactive rather than proactive.

We recommend that the research community takes proactive steps to enhance awareness, training and public engagement regarding synthetic data. Specifically, we recommend developing targeted training modules that equip researchers with a clear understanding of when and how to use synthetic data responsibly, ensuring they can accurately assess its suitability for varied tasks. In addition, we recommend establishing public engagement initiatives to dispel misconceptions about synthetic data.

Final reflections

These four areas of recommendations are interdependent but considered sequentially. Legal and governance clarity enables confidence. Investment in capacity allows organisations to act. Practical tools facilitate consistent and scalable adoption. Training and engagement ensure trust and responsible use.

A phased strategy might take the following steps:

1. **We need governance clarity** so that data creators and TREs know what they can do.
2. **We need funding** so that they have the capacity and expertise to act.
3. **We need tools and standards** so that synthetic data can be implemented consistently.
4. **We need engagement** so that synthetic data is understood, accepted, and used responsibly.

Taken together, these steps would support the use of synthetic data to transition from an experimental tool to a trusted, scalable asset within the UK research infrastructure.

Acknowledgements

The authors would like to thank all the data owners who generously shared their insights and expertise through the survey and case studies. We are also grateful to the TRE professionals who participated in our focus group, contributing invaluable perspectives on synthetic data's practical and governance challenges. Their openness and engagement were instrumental in shaping the findings and recommendations presented in this report.

We would also like to acknowledge our DELIMIT colleagues, Dr Fiona Lugg-Widger and Robert Trubey, for their invaluable support, and we offer special thanks to Emily Oliver from ADR UK for her outstanding guidance and help.

www.ukdataservice.ac.uk

[project page](#)

datasharing@ukdataservice.ac.uk